AD_____

Award Number: W81XWH-05-1-0026

TITLE: Computational Genomics Tools for Copy-Number Fluctuations in Prostate Cancer

PRINCIPAL INVESTIGATOR: Bhubaneswar Mishra, Ph.D.

CONTRACTING ORGANIZATION: New York University
New York, NY 10012-1091

REPORT DATE: November 2005

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-11-2005	Final	1 Nov 2004 – 31 Oct 2005

4. TITLE AND SUBTITLE
Computational Genomics Tools for Copy-Number Fluctuations in Prostate Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-05-1-0026

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Bhubaneswar Mishra, Ph.D.

E-mail: mishra@nyu.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

New York University
New York, NY 10012-1091

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this research our basic goal has been to produce a useful, open-access, large database of lesions in prostate cancer and organize them in terms of segments of aberrant copy numbers for subsequent automated statistical analysis. As a primary example of such analysis, we aimed to enable this database to be easily investigated for enumerating those regions, which harbor genes likely to be causally related to the disease. We hoped to maximize the utility of this database, by optimizing various factors that the design depends upon: cost, availability, efficiency, quality control, and the ease with which it could be studied, navigated and visualized.

15. SUBJECT TERMS
No subject terms provided.

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					USAMRMC
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U	UU	8	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Content:

Introduction……..4
Body……4
Key research accomplishments……4
Reportable outcomes……...4
Conclusions……..5
References……5
Appendices……5

Introduction

In this research our basic goal has been to produce a useful, open-access, large database of lesions in prostate cancer and organize them in terms of segments of aberrant copy numbers for subsequent automated statistical analysis. As a primary example of such analysis, we aimed to enable this database to be easily investigated for enumerating those regions, which harbor genes likely to be causally related to the disease. We hoped to maximize the utility of this database, by optimizing various factors that the design depends upon: cost, availability, efficiency, quality control, and the ease with which it could be studied, navigated and visualized.

Body

Briefly, we have made significant progress in all of these directions:

Key research Accomplishments & Reportable Outcomes

- We have implemented an operational software for oncogenomics (arrayCGH analysis). We subsequently created a significantly improved version of this software ready for the NYU faculties (Ostrer- lab: Mr. Perlman, Ms. Salman) and a number of other collaborators through an online service. This online service integrates the genome view of copy number data with major sources of genome annotation such as NCBI's MapView, KEGG and AmiGO.
- We have used this opportunity to also provide computational training to the researchers in Medical School in using the VALIS platform. We have developed a simple tool for NYU-Medical School labs to browse genomic data with copy number variation as an instructional exercise.
- We have received internal funding from NYU Medical School for one year and plan to release the software through their portals.
- A publication (accessible to biologists and clinical scientists) entitled "A versatile statistical analysis algorithm to detect copy number variation." and describing our initial work under this project was published by PNAS and has attracted major collaborators who are providing us with new data. Four follow-up publications have3 been or are being submitted: (1) A better analysis algorithm for Affymetrix chips (10K, 100K and 500K), (2) Detection of Tumor Suppressor Genes (with exact boundaries) from LOH data, (3) A hyper-parametric model for cancer data and an algorithm based on this model that can handle noisier technologies without biological replicates, and (4) A comparative analysis of data produced by Ostrer-lab.
- The software has been released to the national research community through lab's website and is planned to be released under "Bioconductor" open-access software library.
- Dr. Daruwala et al., with support from this grant, have created "web-spidering" software to aggregate the dispersed information on the Internet about the new Affymetrix CGH chip. The database created from the collected information helps the clinical experimentalists to better interpret the data.
- A computer science graduate student, Ms. Ionita, joined the group about a year ago and has made significant progress towards her PhD thesis research on LOH analysis to detect tumor suppressor genes.
- We have designed algorithms to understand haplotype-copy-number-polymorphisms (HCNP) in a population and in cancer patients. We have designed experiments to perform population studies and a detailed plan to carry out data collection.

Conclusions

In summary, we have made significant and visible progress towards the following three goals:

1. We have developed new statistical methods to improve detection of abnormal lesions, define confidence in the detected lesions, and localize putative genes involved in the cancer.
2. We have created a database with improved statistical significance. It has an enhanced human-computer interface in order that the users (initially, our collaborating teams of scientists and clinicians) can effortlessly maneuver through the data to draw conclusions.
3. Finally, we have created the foundations to build two important "bridges" to future work. The first is a novel statistical algorithm to combine the genomic data with whole genome data for SNP and other markers (for instance indicating LOH). The second is better low-level background correction software that makes the genomic data usable without too many expensive biological replicates. We have now software based on "redescription" to create the capability to easily combine the genomic data with gene-expression data.

References

(1) I. Ionita and R. Daruwala and B. Mishra, "Mapping Tumor Suppressor Genes using Multipoint Statistics from Copy-Number Variation Data," Journal Submission, 2005.

(2) R.-S. Daruwala, A. Rudra, H. Ostrer, R. Lucito, and M. Wigler and B. Mishra, "A Versatile Statistical Analysis Algorithm to Detect Genome Copy Number Variation," (with R.-S. Daruwala, A. Rudra, H. Ostrer, R. Lucito, and M. Wigler), *Proc. National Academy of Science U S A*, **101**(46): 16292-7, 2004.

Appendices

Graduated Ph.D. students.

- *Yi (Joey) Zhou, Genome Evolution, Ph.D. in Biology, 2005.*

Current Ph.D. students.

- *Fang Chen, Genome Evolution, Ph.D. in Biology, Expected Graduation Date: 2005.*
- *Ofer Gill, Algorithms for Genome Alignment, Ph.D. in CS, Expected Graduation Date: 2006.*
- *Matthias Heymann, Evolutionary Processes, Ph.D. in Math, Expected Graduation Date: 2007.*
- *Iuliana Ionita, Haplotype Algorithms, Ph.D. in CS, Expected Graduation Date: 2007.*
- *Venkatesh P. Mysore, Algorithmic Environment for Genomics, Ph.D. in CS, Expected Graduation Date: 2007.*
- *Bing Sun, Physical Models of Genome, Ph.D. in CS, Expected Graduation Date: 2006.*
- *Ilya Rudkevich, Systems Biology, COB program, Mt. Sinai School of Medicine.*

Refereed journal articles that have appeared:

- "A Coherent Framework for Multi-resolution Analysis of Biological Networks with Memory: RAS pathway, Cell Cycle and Immune System," (with P. Barbano, M. Spivak, J. Feng, and M. Antoniotti), Proc. National Academy of Science U S A, 102(18):6245-6250, 2005.

- "On Large-Segmental Duplications in Human Genome and Their Statistical Analysis," (with Y. Zhou), Proc. National Academy of Science U S A, 102(11):4051-4056, 2005.
- "Taming the Complexity of Biochemical Models through Bisimulation and Collapsing: Theory and Practice," (with M. Antoniotti, C.Piazza, A. Policriti and M. Simeoni), Theoretical Computer Science, 325(1): 45-67, 2004.
- "A Versatile Statistical Analysis Algorithm to Detect Genome Copy Number Variation," (with R.-S. Daruwala, A. Rudra, H. Ostrer, R. Lucito, and M. Wigler), Proc. National Academy of Science U S A, 101(46): 16292-7, 2004.
- "Distribution of Short Paired Duplications in Mammalian Genomes." (with E. Thomas et al.), Proc. National Academy of Science U S A, 101(28):10349-10354, 2004.

Chapters in books:

- "Stability of Hybrid Systems and Related Questions from Systems Biology," (In honor of Professor Pravin Varaiya on his 65th birthday), (with C. Piazza), Systems & Control: Foundations & Applications, (Ed. T. Basar), Birkhauser, 2005.
- "Simpathica: A Computational Systems Biology Tool within the Valis Bioinformatics Environment," (with M. Antoniotti, S. Paxia and N. Ugel), Computational Systems Biology, (Ed. E. Eiles and A. Kriete), Elsevier, 2005.

Refereed journal articles accepted for publication

- "Validation of S. pombe sequence assembly by micro-array hybridization," (with J. West, W. Casey, and M. Wigler), Journal of Computational Biology, 2005.
- "Multiple Biological Model Classification: From System Biology to Synthetic Biology," (with M. Antoniotti et al.), Transactions on Computational Systems Biology, 2005.

Research Presentations

- LaserMED Seminar, Center for Catastrophe Preparedness and Response, NYU, NYC, NY, September 23, 2005. “Large Scale Multi-Agent Modeling of Catastrophes: The Brazilian Food-Poisoning Scenario & Beyond”
- Bioinformatics Seminar, Cold Spring Harbor Laboratory, Long Island, NY, September 21, 2005. “Ontology-Based Analysis of Time-Course Gene-Expression Data.”
- Banbury Center Conference on From Markers to Models: Integrating Data to Make Sense of Biologic Systems, Banbury, LI, NY, September 19, 2005. “Computational and Experimental Framework to Understand Disease Pathogenesis.”
- Second International School on Biology, Computation and Information (BCI 2005), Dobbiaco (BZ), Italy, September 15, 2005. “Interpreter of Maladies: Computational and Technological Challenges of Human cancer Genome Project.”
- 8th International Meeting of the Microarray Gene Expression Data Society, MGED 8, Bergen, Norway, September 13, 2005. “Remembrance of Experiments Past: Analyzing Time Course Datasets to Discover Complex Temporal Invariants.”
- Cancer Institute Seminar, University of California at San Diego, SD, CA, August 25, 2005. “Interpreter of Maladies: Computational and Technological Challenges of Human cancer Genome Project.”
- Infosys IT Seminar, Bhubaneswar, Orissa, India, July 21, 2005. “Hitchhiker’s Guide to Bioinformatics.”
- Dabur India Ltd., Ghaziabad, UP, India, July 12, 2005. “Human Cancer Genome Project: India’s Pharmaceutical Industries.”
- National Institute of Immunology., New Delhi, India, July 12, 2005. “Human Cancer Genome Project: Challenges for Bioinformatics.”
- Biotechnology Seminar, Indian Institute of Technology, New Delhi, India, July 12, 2005. “What’s Next? Challenges from Systems Biology.”

- Department of Biotechnology, Ministry of Science & Technology, New Delhi, India, July 11, 2005. “Human Cancer Genome Project: Interpreter of Maladies.”
- 17th Int. Conference on Computer Aided Verification, CAV ’05, Edinburgh, Scotland, UK, July 6, 2005. “What’s Next? Challenges from Systems Biology.”
- Principal Investigators Meeting, NHGRI DNA Sequencing Technology Development Program, Harvard Medical School, Boston, MA, June 23, 2005. “SMASH (Single Molecule Approach to Sequencing by Hybridization).”
- DIMACS Workshop on Detecting and Processing Regularities in High Throughput Biological Data, DIMACS, Rutgers University, NJ, June 20, 2005. “Comparative Genomics: The Lion, the Leopard, the Wolf and the Boar, Why not more?”
- I3P Meeting, Dartmouth Institute for Information Infrastructure Protection (I3P) Consortium Meeting, Puck Building, Wagner School, NYU, NY, June 16 2005. “VALIS Bioinformatics Environment and its Applications to Biosensing.”
- 3rd Annual NYU Cancer Institute Retreat, The Translational Research Program, NYU School of Medicine, Wave Hill, Bronx, NY, June 15, 2005. “Cancer Bioinformatics: Interpreter of Maladies.”
- 2005 Howard Hughes Seminar, Dept. of Biology, NYU, NY, June 14, 2005. “Hitchhiker’s Guide to Bioinformatics.”
- Mathematics Seminar, SUNY, Stony Brook, LI, NY, May 6, 2005. “Systems Biology from the Algebraic Point of View.”
- BioTechnology Seminar, SUNY, Stony Brook, LI, NY, May 6, 2005. “Hooplas, Hypes and Haplotypes: The Joys of Single Molecules.”
- DARPA Biocomp Meeting, Arlington, VA, May 3, 2005. “Computational Models,Reasoning Tools, & Applications (Part III).”
- Applied Math Seminar, Courant Institute, New York, NY, April 22, 2005. “The Lion, the Leopard, the Wolf and the Boar, Why not more? Algorithms for Comparative Genomics.”
- ITL Seminar Series (MEL & CSTL), National Institute of Standards and Technology, Gaithersburg, MD, April 15, 2005. “Interpreter of Maladies: Computational Approaches to Biomedical Problems.”
- Workshop on Computable Semantics for Complex Biological Systems, Arlington, VA, March 3, 2005. “Time-Course Analysis of Host-Pathogen Interaction with GOALIE.”
- Distinguished Lecture Series, University of Maryland, College Park, MA, November 17, 2004. “Cell Talk.”
- Distinguished Seminar Series, Drexel University, Philadelphia, PA, October 15, 2004. “Cell Talk.”
- DARPA Biocomp Meeting, Vienna, VA, October 12, 2004. “Computational Models, Reasoning Tools, & Applications (Part II).”
- Minisymposium on Microarray and Bioinfomatics, Temple University, Philadelphia, PA, September 29, 2004. “Identifying Differentially Expressed Genes via Multiscale Geometric Analysis.”
- BioConcur 04, The Royal Society, London, UK, August 28, 2004.
- Biologically Motivated Problems in Statistics, STATPHYS 22, Bangalore, India, July 6, 2004. “Genome Evolution by Substitutions, Duplications and Deletions.”
- Bioinformatics Lecture, Regeneron Pharmaceuticals, Inc., Tarrytown, NY, June 16, 2004. “BAC to the Future.”
- Biogeometry Workshop, Symposium on Computational Geometry, Brooklyn, NY, June 12, 2004. “Cell Talk.”
- Plenary Speaker, International Conference on Complex Systems, Boston, MA, May 20, 2004. “Cell Talk.”
- Bioinformatics Program Seminar, Boston University, Boston, MA, May 20, 2004. “Cell Talk.”
- ECE/CS Distinguished Lecture, Carnegie-Mellon University, Pittsburgh, PA, April 30, 2004. “Cell Talk.”

- Demerec In-house Seminar, Cold Spring Harbor Laboratory, Long Island, NY, April 7, 2004. “”Valis, Simpathica and NYU MAD: Computational and Systems Biology Tools and their Applications.”
- Systems Biology Seminar, Harvard Medical School, Harvard University, Boston, MA, March 15, 2004. “Cell Talk.”
- Bioinformatics Seminar, Cold Spring Harbor Laboratory, Long Island, NY, March 10, 2004. “Identifying Differentially Expressed Genes via Multiscale Geometric Analysis.”
- Scientific Horizons Seminar, SAC Capital Advisors, LLC, New York, NY, February 12, 2004. “Cell Talk.”
- Department of Computer Science, Dartmouth College, Hanover, NH, February 10, 2004. “Cell Talk.”
- DARPA Biocomp Meeting, Adelphi, MD, February 3, 2004. “Computational Models, Reasoning Tools, & Applications to Apoptosis.”